# Cross-Lingual Information Retrieval in Tetun for Ad-Hoc Search

Altedio Araújo[1], Gabriel de Jesus[2](✉), and Sérgio Nunes[2]

[1] School of Management and Technology, Polytechnic of Porto (ESTG), Felgueiras, Portugal
8220759@estg.ipp.pt

[2] INESC TEC/Faculty of Engineering, University of Porto (FEUP), Porto, Portugal
gabriel.jesus@inesctec.pt, sergio.nunes@fe.up.pt

**Abstract.** Developing information retrieval (IR) systems that enable access across multiple languages is crucial in multilingual contexts. In Timor-Leste, where Tetun, Portuguese, English, and Indonesian are official and working languages, no cross-lingual information retrieval (CLIR) solutions currently exist to support information access across these languages. This study addresses that gap by investigating CLIR approaches tailored to the linguistic landscape of Timor-Leste. Leveraging an existing monolingual Tetun document collection and ad-hoc text retrieval baselines, we explore the feasibility of CLIR for Tetun. Queries were manually translated into Portuguese, English, and Indonesian to create a multilingual query set. These were then automatically translated back into Tetun using Google Translate and several large language models, and used to retrieve documents in Tetun. Results show that Google Translate is the most reliable tool for Tetun CLIR overall, and the Hiemstra LM consistently outperforms BM25 and DFR BM25 in cross-lingual retrieval performance. However, overall effectiveness remains up to 26.95% points lower than that of the monolingual baseline, underscoring the limitations of current translation tools and the challenges of developing an effective CLIR for Tetun. Despite these challenges, this work establishes the first CLIR baseline for Tetun ad-hoc text retrieval, providing a foundation for future research in this under-resourced setting.

**Keywords:** Cross-lingual information retrieval · Tetun · Ad-hoc text retrieval · Low-resource language

## 1 Introduction

Cross-lingual information retrieval (CLIR) systems are essential for enabling users to submit queries in one language and retrieve relevant documents written in another. Such systems are particularly valuable in linguistically diverse contexts, such as Timor-Leste. As a multilingual country, Timor-Leste designates Tetun and Portuguese as its official languages, and English and Indonesian as working languages [27], with over 30 dialects spoken across the territory [7].

J. Valente de Oliveira et al. (Eds.): EPIA 2025, LNAI 16122, pp. 262–275, 2026.
https://doi.org/10.1007/978-3-032-05179-0_20

Within this complex linguistic landscape, CLIR becomes a crucial tool for facilitating access to information across these languages.

A classical CLIR system typically consists of a translation step followed by monolingual retrieval [29]. Two primary translation strategies are commonly employed: **query translation**, where the user's query is translated into the language of the documents, and **document translation**, where documents are translated into the language of the query [18,30]. Query translation is generally more efficient and conducive to rapid experimentation, whereas document translation can be computationally expensive and time-intensive, especially when translating an entire corpus [16,24].

Tetun, a dialect that became one of the official languages following the country's restoration of independence from Indonesia in 2002, remains under-resourced, with few algorithms and tools available to support the language. Notably, the first initiative focused on Tetun text information retrieval (IR) began in 2022 [7,8], leading to the creation of foundational resources for IR and natural language processing (NLP) [10–12], including a test collection and baselines for the ad-hoc text retrieval task [9].

Building on these efforts, and given the lack of existing CLIR for Tetun, this study explores the effectiveness of CLIR for retrieving Tetun documents using queries in Portuguese, English, and Indonesian. Results from this study are compared against existing monolingual Tetun ad-hoc text retrieval baselines [9] to evaluate the performance of CLIR for Tetun. This study is guided by the following research questions (RQs):

> **RQ1:** How does the quality of query translation influence the effectiveness of Tetun ad-hoc text retrieval?
>
> **RQ2:** How does the effectiveness of CLIR compare to that of monolingual retrieval in Tetun?

To address these RQs, we first manually translated the original Tetun queries from *Labadain-Avaliadór* [12] into Portuguese, English, and Indonesian, resulting in a multilingual query set. These translated queries served as input to the CLIR system for retrieving documents in Tetun. To enable retrieval, the multilingual queries were automatically translated back into Tetun using both Google Translate and large language models (LLMs) before being submitted to the Tetun monolingual retrieval system.

To respond to RQ1, we evaluate the quality of the translated queries using the BLEU metric [20], with the original Tetun queries serving as reference texts. This RQ aims to examine whether translation quality directly influences the effectiveness of the CLIR system. For RQ2, we compare the retrieval performance of the CLIR system to the established monolingual baselines, evaluating effectiveness using P@10, MAP@10, and NDCG@10 for each source language.

The findings provide initial insights into the feasibility of CLIR for Tetun, demonstrating the impact of translation quality on retrieval effectiveness. This highlights the potential of machine translation-based query methods to enhance the performance of Tetun ad-hoc text retrieval.

The remainder of this paper is organized as follows. Section 2 reviews the background and related work. Section 3 presents the methodology, followed by the CLIR experimental setup in Sect. 4. Section 5 reports the evaluation and results, which are further discussed in Sect. 6. Finally, Sect. 7 summarizes the main conclusion and outlines directions for future work.

## 2 Background and Related Work

Tetun is an Austronesian language spoken in Timor-Leste, an island country in Southeast Asia. Historically used as a lingua franca in trade and religious activities, Tetun became one of Timor-Leste's official languages when the country restored its independence in May 2002. It is a low-resource language spoken by approximately 79% of the country's 1.18 million population, according to the 2015 census [6,10]. The development of Tetun IR and NLP algorithms and tools began in 2022 [8], leading to the creation of key resources, including a text dataset [11] and a test collection with baselines for the ad-hoc text retrieval task [9]. These tools and resources provide the essential foundation for investigating CLIR in Tetun.

CLIR has been extensively studied, with several works exploring its application to low-resource languages (LRLs). Adeyemi et al. [2] conducted CLIR experiments covering several African languages, including Hausa, Somali, Swahili, and Yoruba, using English queries. Their study investigated both manual and automatic query translation from English into the target languages, and automatic translation of documents into English. For query translation, they used Google Translate, and reported that human-translated queries retrieved using BM25 outperformed automatically translated ones by 5.78% points in NDCG@20. In the CLEF 2009 track [4], the use of Google Translate for query translation significantly improved cross-lingual retrieval performance. Similarly, a study focusing on English to Chinese retrieval showed that translating queries using Google Translate could achieve performance comparable to monolingual baselines, particularly when queries are longer [28].

Several studies have explored the use of LLMs for query translation in CLIR. Adeyemi et al. [1] investigated the effectiveness of LLMs for query translation by adopting zero-shot prompting with GPT-3.5 and GPT-4 to translate English queries into Hausa, Somali, Swahili, and Yoruba. The translation quality was evaluated using BLEU [20], and GPT-4 achieved the highest scores for all languages. Similarly, Valentini et al. [26] translated both queries and documents from English to French using zero-shot prompting with GPT-4o-mini and LLaMA-3.2, evaluating translation quality using several metrics, including BLEU. For document translation, GPT-4o-mini achieved a higher BLEU score than LLaMA-3.2 and was subsequently used for query translation. For retrieval, BM25 was used, and document translation with GPT-4o-mini achieved better performance than LLaMA-3.2 in terms of Recall@100 and NDCG@10.

Pecina et al. [21] investigated CLIR in the medical domain, focusing on improving query translation quality and examining its impact on retrieval effectiveness. They used Moses [13], a phrase-based statistical machine translation system trained on medical domain data, and compared its performance with Google Translate and Microsoft Bing Translator. Retrieval was conducted in English using the CLEF eHealth 2013 collection, while the original queries (written in Czech, German, and French) were translated into English. In terms of translation quality, Moses outperformed both Google Translate and Microsoft Bing Translator. For retrieval, Moses maintained superior performance compared to Bing Translator in Czech to English translation at P@10, but it did not consistently outperform Google Translate across all input languages. Overall, the authors reported that higher BLEU scores in translation did not lead to improved retrieval effectiveness.

Likewise, Lignos et al. [15] examined the relationship between machine translation quality and retrieval effectiveness in CLIR for Czech and German, focusing on document translation. In their study, queries were written in English, and documents in Czech and German were translated into English. In line with the findings of Pecina et al. [21], higher BLEU scores in document translation did not enhance retrieval performance.

Building on this foundation, and given recent advancements in LLMs and the inclusion of Tetun to Google Translate,[1] new opportunities have emerged to explore the applicability of these approaches to Tetun. In this study, we evaluate the application of Google Translate alongside several LLMs to assess their effectiveness in translating queries into Tetun and compare their performance against established monolingual Tetun IR baselines.

## 3 Methodology

To investigate CLIR for Tetun, we use *Labadain-Avaliadór* [12], a Tetun test collection for the ad-hoc text retrieval task, which contains 33,500 documents, 59 queries, and 5,900 relevance judgments (qrels) [9]. The study began by manually translating Tetun queries into Portuguese, English, and Indonesian. The translations were conducted by three native Tetun-speaking volunteer students, each responsible for a single target language in which they are fluent. The process followed established guidelines, including specific instructions to preserve abbreviations and named entities, resulting in a multilingual query set.

These multilingual queries are subsequently used as input for retrieval. During the query processing phase, the input queries are automatically translated back into Tetun using Google Translate and LLMs, including GPT-4.5, DeepSeek-Chat, Claude 3.7 Sonnet, and LLaMA 3.3 70B. To assess translation quality, the outputs are evaluated using the BLEU metric [20], with the original Tetun queries serving as reference texts.

The translated queries are then used for retrieval. To enable performance comparison, we adopt the best-performing retrieval strategy previously reported in monolingual Tetun ad-hoc text retrieval. The retrieval and ranking models used in the experiments are BM25 [23], DFR BM25 [3], and the Hiemstra language model (LM) [22]. These models are evaluated using Precision, MAP, and

[1] https://blog.google/products/translate/google-translate-new-languages-2024/.

NDCG at rank ten ($k = 10$) for each input query in English, Portuguese, and Indonesian. Results from CLIR are compared to the monolingual Tetun ad-hoc text retrieval baselines [9] to assess their relative performance.

# 4 Experimental Setup

The experiments are organized into two parts: query translation and cross-lingual retrieval. Three query translation strategies were evaluated: Google Translate, LLMs with zero-shot prompting, and LLMs with few-shot prompting using one example. The experimental setup for each configuration is described below.

## 4.1 Query Translations

Query translation using Google Translate follows the official documentation and default settings of the basic translation endpoint[2] provided by the Google Translate API. For LLM-based translations, the same prompt is used across all models, with the temperature set to zero to ensure deterministic outputs.

The prompt used to instruct the LLMs was compiled based on best practices for prompting translation models [1,5,25].[3] Details of the zero-shot prompting are presented in Prompt 4.1.

Prompt 4.1. Zero-shot prompt used for query translation with LLMs

> You are an expert translator. Your task is to translate the following user search query from *{source_lang}* into Tetun.
>
> Query: *{query}*
>
> Translate the query accurately, using the official grammar of Tetun and correct vocabulary. Return only the translation and do not include any additional text.

For few-shot prompting, we used the same base prompt as in the zero-shot setting, with one example query and its corresponding translation in each target language included in the prompt for each LLM call. The additional instruction appended to the base prompt is shown below:

> Refer to the example below to guide your translation:
>
> *{source_lang}* query: *{example_query}*<br>
> Tetun translation: *{example_translation}*

[2] https://cloud.google.com/translate/docs/reference/rest/v2/translate.

[3] https://docs.anthropic.com/en/resources/prompt-library/polyglot-superpowers.

### 4.2 Cross-Lingual Retrieval

For retrieval, we employed the monolingual Tetun ad-hoc text retrieval framework adopted by de Jesus and Nunes [9]. We used PyTerrier [17], a Python API for the Terrier IR platform [19], for indexing, retrieval, and ranking, using the default settings for each retrieval model. This setup indexes only the document titles and applies uniform preprocessing to both queries and titles, including lowercasing and the removal of punctuation, special characters, apostrophes, and hyphens. All queries translated from Portuguese, English, and Indonesian into Tetun, using each of the translation techniques described in Subsect. 4.1, were employed for retrieval and evaluated using *Labadain-Avaliadór* [12].

## 5 Evaluation and Results

The evaluation focuses on the quality of query translations and the effectiveness of CLIR compared to monolingual Tetun ad-hoc text retrieval. Translation quality was measured using BLEU [20], while retrieval effectiveness was assessed at P@10, MAP@10, and NDCG@10. The evaluation results are shown in Table 1.

**Table 1.** BLEU scores for translation quality from Portuguese (Pt), English (En), and Indonesian (Id) into Tetun (Tdt).

| Model | Zero-Shot | | | Few-Shot | | |
|---|---|---|---|---|---|---|
| | **PT-Tdt** | **En-Tdt** | **Id-Tdt** | **Pt-Tdt** | **En-Tdt** | **Id-Tdt** |
| Google Translate | **0.4659** | 0.4530 | 0.5636 | - | - | - |
| GPT-4.5-preview | 0.4156 | **0.4882** | 0.6038 | 0.4215 | 0.4156 | 0.4616 |
| DeepSeek-Chat | 0.3655 | 0.3762 | 0.5091 | 0.4634 | 0.4183 | 0.4994 |
| Claude-3.7 Sonnet | 0.4602 | 0.4649 | **0.6256** | **0.5378** | **0.5412** | **0.6425** |
| LLaMA-3.3 70B | 0.3566 | 0.3184 | 0.4172 | 0.3704 | 0.3122 | 0.4033 |

As observed, under the zero-shot setting, BLEU scores varied across source languages and models. Google Translate achieved the highest score for Portuguese-to-Tetun translation, GPT-4.5 performed best for English-to-Tetun, and Claude 3.7 Sonnet achieved the highest score for Indonesian-to-Tetun. In the few-shot setting, Claude 3.7 Sonnet recorded the highest BLEU scores for all languages.

When comparing BLEU scores from few-shot and zero-shot LLM-based translations, the following results are observed: (1) **Portuguese-to-Tetun:** translation quality improved across all models in the few-shot setting, with gains of up to +9.79% points; (2) **English-to-Tetun:** results varied by model, performance dropped for GPT-4.5-preview and LLaMA-3.3 70B by up to -7.25% points, while DeepSeek-Chat and Claude 3.7 Sonnet showed improvements of up to +7.63% points; and (3) **Indonesian-to-Tetun:** only Claude 3.7 Sonnet showed an improvement (+1.69% points), while the other models experienced performance drops of up to -14.22% points.

Regarding the evaluation of retrieval performance, results using zero-shot query translations are presented in Table 2. The findings show that translating input queries with Google Translate, followed by retrieval and ranking using the Hiemstra LM, generally outperformed the LLM-based approaches. An exception was observed in Portuguese-to-Tetun translation, where LLMs achieved higher scores at P@10 and MAP@10. Nevertheless, all CLIR results remained significantly below the performance of the monolingual Tetun IR baselines, by up to -26.95% points for Portuguese-to-Tetun translation, -25.93 for English-to-Tetun, and -23.22 for Indonesian-to-Tetun.

Evaluating the retrieval performance for few-shot prompting with one example per language, Claude 3.7 Sonnet outperformed all other models across all languages and metrics, as shown in Table 3. Compared to the zero-shot results, providing one example per LLM call for Portuguese-to-Tetun translation improved retrieval performance by up to +0.93% points. In contrast, for English-to-Tetun, the inclusion of an example resulted in a performance drop of up to -2.21% points. For Indonesian-to-Tetun, the differences were marginal: a slight performance drop in P@10 and NDCG@10 (up to -0.63% points), while MAP@10 showed a small gain of +0.29% points.

## 6 Discussion

The effectiveness of CLIR in Tetun varies depending on the translation model and the prompting strategy applied to LLMs. In the zero-shot setting, Google Translate consistently demonstrated stronger retrieval performance overall, particularly for Indonesian-to-Tetun query translation. When paired with the Hiemstra LM, it achieved the highest scores among all zero-shot configurations evaluated. The effectiveness of Google Translate for query translation in CLIR for Tetun aligns with findings from the CLEF 2009 track [4], where its use substantially improved CLIR performance. Likewise, studies on English-to-Chinese CLIR reported retrieval results only marginally lower than monolingual ones when Google Translate was employed for query translation [28].

Under the few-shot configuration, Claude 3.7 Sonnet showed the strongest retrieval performance for all languages, consistent with its strong zero-shot performance in Portuguese-to-Tetun translation, particularly at P@10 and MAP@10. For Portuguese-to-Tetun query translation, few-shot prompting led to a marginal improvement of up to +0.93% points compared to the zero-shot setting. However, for English and Indonesian queries, even when the example was provided, LLM-based translations did not surpass the retrieval performance using Google Translate. These results underscore the effectiveness of Google Translate as a reliable translation tool for CLIR in Tetun under zero-shot settings.

**Table 2.** Effectiveness of CLIR using zero-shot translated queries compared to the monolingual baseline. Values in **bold** indicate the best CLIR performance for each source language. The baseline, highlighted in orange, indicates the overall best performance across all retrieval metrics.

| **Translation** | **Language** | **Model** | **P@10** | **MAP@10** | **NDCG@10** |
|---|---|---|---|---|---|
| Baseline | Tetun | DFR BM25 | 0.8390 | 0.2804 | 0.7356 |
| Google Translate | Portuguese - Tetun | BM25 | 0.5559 | 0.1659 | 0.4934 |
| | | DFR BM25 | 0.5576 | 0.1661 | 0.4931 |
| | | Hiemstra LM | 0.5576 | 0.1706 | **0.4914** |
| | English - Tetun | BM25 | 0.5712 | 0.1633 | 0.4937 |
| | | DFR BM25 | 0.5695 | 0.1634 | 0.4907 |
| | | Hiemstra LM | **0.5797** | **0.1694** | **0.4971** |
| | Indonesian - Tetun | BM25 | 0.6000 | 0.1861 | 0.5216 |
| | | DFR BM25 | 0.6017 | 0.1865 | 0.5226 |
| | | Hiemstra LM | **0.6068** | **0.1893** | **0.5237** |
| GPT-4.5-preview | Portuguese - Tetun | BM25 | 0.4864 | 0.1472 | 0.4230 |
| | | DFR BM25 | 0.4881 | 0.1474 | 0.4235 |
| | | Hiemstra LM | 0.4983 | 0.1565 | 0.4257 |
| | English - Tetun | BM25 | 0.4966 | 0.1463 | 0.4161 |
| | | DFR BM25 | 0.4966 | 0.1467 | 0.4155 |
| | | Hiemstra LM | 0.4847 | 0.1497 | 0.4149 |
| | Indonesian - Tetun | BM25 | 0.5407 | 0.1646 | 0.4576 |
| | | DFR BM25 | 0.5424 | 0.1654 | 0.4582 |
| | | Hiemstra LM | 0.5339 | 0.1614 | 0.4495 |
| DeepSeek-Chat | Portuguese - Tetun | BM25 | 0.4407 | 0.1246 | 0.3739 |
| | | DFR BM25 | 0.4407 | 0.1247 | 0.3734 |
| | | Hiemstra LM | 0.4441 | 0.1262 | 0.3754 |
| | English - Tetun | BM25 | 0.4373 | 0.1329 | 0.3695 |
| | | DFR BM25 | 0.4373 | 0.1337 | 0.3686 |
| | | Hiemstra LM | 0.4407 | 0.1364 | 0.3719 |
| | Indonesian - Tetun | BM25 | 0.5000 | 0.1490 | 0.4232 |
| | | DFR BM25 | 0.5000 | 0.1491 | 0.4222 |
| | | Hiemstra LM | 0.4932 | 0.1439 | 0.4139 |
| Claude-3.7 Sonnet | Portuguese - Tetun | BM25 | 0.5475 | 0.1678 | 0.4763 |
| | | DFR BM25 | 0.5508 | 0.1689 | 0.4774 |
| | | Hiemstra LM | **0.5695** | **0.1779** | 0.4912 |
| | English - Tetun | BM25 | 0.5678 | 0.1571 | 0.4656 |
| | | DFR BM25 | 0.5627 | 0.1563 | 0.4624 |
| | | Hiemstra LM | 0.5610 | 0.1625 | 0.4717 |
| | Indonesian - Tetun | BM25 | 0.5763 | 0.1868 | 0.4910 |
| | | DFR BM25 | 0.5797 | 0.1879 | 0.4920 |
| | | Hiemstra LM | 0.5763 | 0.1833 | 0.4890 |
| LLaMA-3.3 70B | Portuguese - Tetun | BM25 | 0.3831 | 0.1060 | 0.3197 |
| | | DFR BM25 | 0.3814 | 0.1060 | 0.3185 |
| | | Hiemstra LM | 0.3729 | 0.0990 | 0.3014 |
| | English - Tetun | BM25 | 0.3508 | 0.0896 | 0.2896 |
| | | DFR BM25 | 0.3508 | 0.0899 | 0.2887 |
| | | Hiemstra LM | 0.3525 | 0.0921 | 0.2936 |
| | Indonesian - Tetun | BM25 | 0.3763 | 0.1149 | 0.3045 |
| | | DFR BM25 | 0.3763 | 0.1145 | 0.3040 |
| | | Hiemstra LM | 0.3780 | 0.1139 | 0.3043 |

**Table 3.** Effectiveness of CLIR using few-shot translated queries compared to the monolingual baseline. Values in **bold** indicate the best CLIR performance for each source language. The baseline, highlighted in orange, indicates the overall best performance across all retrieval metrics.

| Translation | Language | Model | P@10 | MAP@10 | NDCG@10 |
|---|---|---|---|---|---|
| Baseline | Tetun | DFR BM25 | 0.8390 | 0.2804 | 0.7356 |
| GPT-4.5-preview | Portuguese - Tetun | BM25 | 0.4373 | 0.1282 | 0.3755 |
| | | DFR BM25 | 0.4373 | 0.1283 | 0.3742 |
| | | Hiemstra LM | 0.4458 | 0.1335 | 0.3775 |
| | English - Tetun | BM25 | 0.4508 | 0.1291 | 0.3696 |
| | | DFR BM25 | 0.4492 | 0.1287 | 0.3684 |
| | | Hiemstra LM | 0.4542 | 0.1313 | 0.3738 |
| | Indonesian - Tetun | BM25 | 0.4847 | 0.1406 | 0.4050 |
| | | DFR BM25 | 0.4847 | 0.1406 | 0.4050 |
| | | Hiemstra LM | 0.4627 | 0.1323 | 0.3847 |
| DeepSeek-Chat | Portuguese - Tetun | BM25 | 0.4915 | 0.1509 | 0.4253 |
| | | DFR BM25 | 0.4932 | 0.1513 | 0.4252 |
| | | Hiemstra LM | 0.5000 | 0.1527 | 0.4321 |
| | English - Tetun | BM25 | 0.4746 | 0.1354 | 0.3921 |
| | | DFR BM25 | 0.4729 | 0.1353 | 0.3911 |
| | | Hiemstra LM | 0.4576 | 0.1330 | 0.3850 |
| | Indonesian - Tetun | BM25 | 0.5237 | 0.1648 | 0.4454 |
| | | DFR BM25 | 0.5237 | 0.1648 | 0.4443 |
| | | Hiemstra LM | 0.5220 | 0.1615 | 0.4376 |
| Claude-3.7 Sonnet | Portuguese - Tetun | BM25 | 0.5661 | 0.1692 | 0.4987 |
| | | DFR BM25 | 0.5695 | 0.1713 | 0.5011 |
| | | Hiemstra LM | **0.5797** | **0.1787** | **0.5091** |
| | English - Tetun | BM25 | 0.5576 | 0.1612 | 0.4702 |
| | | DFR BM25 | 0.5593 | 0.1628 | 0.4700 |
| | | Hiemstra LM | **0.5661** | **0.1685** | **0.4750** |
| | Indonesian - Tetun | BM25 | 0.6034 | 0.1914 | 0.5168 |
| | | DFR BM25 | **0.6051** | **0.1922** | **0.5174** |
| | | Hiemstra LM | 0.5949 | 0.1859 | 0.5070 |
| LLaMA-3.3 70B | Portuguese - Tetun | BM25 | 0.3729 | 0.1096 | 0.2940 |
| | | DFR BM25 | 0.3729 | 0.1103 | 0.2934 |
| | | Hiemstra LM | 0.3661 | 0.1068 | 0.2869 |
| | English - Tetun | BM25 | 0.3271 | 0.0769 | 0.2623 |
| | | DFR BM25 | 0.3271 | 0.0768 | 0.2615 |
| | | Hiemstra LM | 0.3254 | 0.0779 | 0.2643 |
| | Indonesian - Tetun | BM25 | 0.3729 | 0.1096 | 0.2940 |
| | | DFR BM25 | 0.3729 | 0.1103 | 0.2934 |
| | | Hiemstra LM | 0.3661 | 0.1068 | 0.2869 |

When examining the relationship between translation quality and retrieval effectiveness, we observed findings consistent with those reported in the literature. For instance, although Google Translate achieved the highest BLEU score for Portuguese-to-Tetun translation (see Table 1), it was outperformed by Claude 3.7 Sonnet in retrieval effectiveness, as shown in Table 2. Similarly, while GPT-4.5-preview achieved the highest BLEU score for English-to-Tetun translation, it yielded lower retrieval performance compared to Google Translate. One possible explanation is that retrieval effectiveness depends not only on translation quality but also on how well the translated queries align with document representations and how effectively the retrieval and ranking models process those queries. Factors such as vocabulary overlap and model-specific ranking behavior may influence retrieval outcomes in ways that BLEU scores do not capture. These findings address RQ1, suggesting that higher translation quality—as measured by BLEU—does not always lead to improved retrieval performance in CLIR. This observation aligns with prior studies by Pecina et al. [21] and Lignos et al. [15], both of which reported similar discrepancies between translation quality and retrieval performance across various domains and language pairs.

When comparing CLIR performance with the monolingual Tetun baselines, addressing RQ2, results remain significantly lower than those achieved by monolingual retrieval. This gap may be explained by the limitations in the translation models' capabilities. Notably, Google Translate only recently added support for Tetun, and the limited number of Tetun documents on the web as of 2023 [10,14] further constrains the ability of LLMs to process the language effectively.

**Limitations and Potential Biases**: The volunteer translators are students, and not all have formal training in linguistics, which may affect the overall quality of the human translations. Additionally, no systematic validation of translation quality was conducted, potentially introducing errors or inconsistencies in the translated queries. Another limitation is the short length of the queries (between three and five words), which provides limited context and may hinder accurate translation, particularly in linguistically ambiguous cases. Furthermore, the translators' backgrounds, language preferences, and interpretation styles may introduce bias into how certain terms or phrases are translated across languages.

## 7 Conclusions and Future Work

This study explored the feasibility and effectiveness of CLIR for Tetun, an under-resourced language, by automatically translating queries from Portuguese, English, and Indonesian using both Google Translate and LLMs. The findings reveal several key insights.

First, despite advances in multilingual technologies, CLIR performance for Tetun remains significantly below that of monolingual retrieval. This highlights the ongoing challenges of supporting LRLs in information access tasks. Second, while LLMs show promise in handling a variety of tasks, including translation, their effectiveness is highly dependent on the source language and their familiarity with or knowledge of the target language. In some cases, traditional machine translation tools like Google Translate prove more reliable, particularly when newly supporting a language like Tetun.

Most critically, our study reveals that translation quality—as measured by standard metrics such as BLEU—does not correlate with retrieval effectiveness. This finding aligns with prior research [15,21] and suggests the need to move beyond translation-centric metrics when evaluating CLIR systems. Retrieval success depends not only on linguistic translation accuracy but also on how well translated queries align with document representations and the retrieval models.

In future work, we aim to explore more effective CLIR systems for languages like Tetun by going beyond improvements in translation techniques. Our focus will include expanding Tetun's digital footprint and rethinking retrieval strategies to better reflect its linguistic characteristics and resource constraints. We will also employ other translation evaluation metrics, such as ROUGE and METEOR, to more accurately assess translation quality and examine its correlation with retrieval effectiveness.

**Acknowledgments.** We thank Luisa Carlos, Zeferina da Graça, and Lara da Costa for translating the Tetun queries into English, Portuguese, and Indonesian, respectively. This work is financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, under the support UID/50014/2023 (DOI: 10.54499/UID/50014/2023). Gabriel de Jesus is supported by the Ph.D. studentship grant SFRH/BD/151437/2021, also from FCT (DOI: 10.54499/SFRH/BD/151437/2021).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Adeyemi, M., Oladipo, A., Pradeep, R., Lin, J.: Zero-shot cross-lingual reranking with large language models for low-resource languages. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, 11–16 August 2024, pp. 650–656. Association for Computational Linguistics (2024). https://aclanthology.org/2024.acl-short.59
2. Adeyemi, M., et al.: CIRAL: a test collection for CLIR evaluations in African languages. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, 14–18 July 2024, pp. 293–302. ACM (2024). https://doi.org/10.1145/3626772.3657884
3. Amati, G., van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. **20**(4), 357–389 (2002). https://doi.org/10.1145/582415.582416
4. Ferro, N., Peters, C.: CLEF 2009 ad hoc track overview: TEL and persian tasks. In: Peters, C., et al. (eds.) Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, 30 September– 2 October 2009, Revised Selected Papers. Lecture Notes in Computer Science, vol. 6241, pp. 13–35. Springer (2009). https://www.dei.unipd.it/~ferro/papers/2010/CLEF2009Proc-adhoc.pdf

5. He, S.: Prompting chatgpt for translation: a comparative analysis of translation brief and persona prompts. In: Scarton, C., et al. (eds.) Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), EAMT 2024, Sheffield, UK, 24–27 June 2024, pp. 316–326. European Association for Machine Translation (EAMT) (2024). https://aclanthology.org/2024.eamt-1.27
6. Instituto Nacional de Estatística Timor-Leste (INETL): Census 2015 priority table population by language (2015). https://inetl-ip.gov.tl/2023/03/09/census-2015-priority-table-population-by-language/. Accessed 17 Mar 2024
7. de Jesus, G.: Text information retrieval in Tetun. In: Kamps, J., et al. (eds.) Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, 2–6 April 2023, Proceedings, Part III. Lecture Notes in Computer Science, vol. 13982, pp. 429–435. Springer (2023). https://doi.org/10.1007/978-3-031-28241-6_48
8. de Jesus, G.: Text information retrieval in Tetun: a preliminary study. CoRR (2024). https://doi.org/10.48550/ARXIV.2406.07331
9. de Jesus, G., Nunes, S.: Establishing a foundation for Tetun text ad-hoc retrieval: indexing, stemming, retrieval, and ranking. CoRR (2024). https://doi.org/10.48550/ARXIV.2412.11758
10. de Jesus, G., Nunes, S.: Labadain-30k+: a monolingual Tetun document-level audited dataset. In: Melero, M., Sakti, S., Soria, C. (eds.) Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, pp. 177–188. ELRA and ICCL, Torino, Italia (2024). https://aclanthology.org/2024.sigul-1.22/
11. de Jesus, G., Nunes, S.: Labadain-30k+: a monolingual Tetun document-level audited dataset [dataset]. INESC TEC (2024). https://doi.org/10.25747/YDWR-N696
12. de Jesus, G., Nunes, S.: Labadain-Avaliadór: A Test Collection for Tetun Ad-hoc Text Retrieval Task [dataset]. INESC TEC (2025). https://doi.org/10.25747/2k6s-e518
13. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 23–30 June 2007, Prague, Czech Republic. The Association for Computational Linguistics (2007). https://aclanthology.org/P07-2045/
14. Kudugunta, S., et al.: MADLAD-400: a multilingual and document-level large audited dataset. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, 10–16 December 2023 (2023). http://papers.nips.cc/paper_files/paper/2023/hash/d49042a5d49818711c401d34172f9900-Abstract-Datasets_and_Benchmarks.html
15. Lignos, C., Cohen, D., Lien, Y., Mehta, P., Croft, W.B., Miller, S.: The challenges of optimizing machine translation for low resource cross-language information retrieval. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, 3–7 November 2019, pp. 3495–3500. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/D19-1353
16. Lin, J., et al.: Simple yet effective neural ranking and reranking baselines for cross-lingual information retrieval. CoRR (2023). https://doi.org/10.48550/ARXIV.2304.01019

17. Macdonald, C., Tonellotto, N.: Declarative experimentation in information retrieval using pyterrier. In: Balog, K., Setty, V., Lioma, C., Liu, Y., Zhang, M., Berberich, K. (eds.) ICTIR 2020: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, 14–17 September 2020, pp. 161–168. ACM (2020). https://doi.org/10.1145/3409256.3409829
18. Ogundepo, O., Zhang, X., Sun, S., Duh, K., Lin, J.: Africlirmatrix: enabling cross-lingual information retrieval for African languages. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022, pp. 8721–8728. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.EMNLP-MAIN.597
19. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: Losada, D.E., Fernández-Luna, J.M. (eds.) Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, 21–23 March 2005, Proceedings. Lecture Notes in Computer Science, vol. 3408, pp. 517–519. Springer (2005). https://doi.org/10.1007/978-3-540-31865-1_37
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 6–12 July 2002, Philadelphia, PA, USA, pp. 311–318. ACL (2002). https://doi.org/10.3115/1073083.1073135
21. Pecina, P., et al.: Adaptation of machine translation for multilingual information retrieval in the medical domain. Artif. Intell. Med. **61**(3), 165–185 (2014). https://doi.org/10.1016/J.ARTMED.2014.01.004
22. Robertson, S., Hiemstra, D.: Language models and probability of relevance. In: Proceedings of the Workshop on Language Modeling and Information Retrieval, pp. 21–25. Carnegie Mellon University, United States (2001)
23. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009). https://doi.org/10.1561/1500000019
24. Saleh, S., Pecina, P.: Document translation vs. query translation for cross-lingual information retrieval in the medical domain. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020, pp. 6849–6860. Association for Computational Linguistics (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.613
25. Tang, L., Qin, J., Ye, W., Tan, H., Yang, Z.: Adaptive few-shot prompting for machine translation with pre-trained language models. In: Walsh, T., Shah, J., Kolter, Z. (eds.) AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, 25 February– 4 March 2025, Philadelphia, PA, USA, pp. 25255–25263. AAAI Press (2025). https://doi.org/10.1609/AAAI.V39I24.34712
26. Valentini, F., Kozlowski, D., Larivière, V.: Clirudit: cross-lingual information retrieval of scientific documents (2025). https://arxiv.org/abs/2504.16264
27. Vasconcelos, P.C.B.D., et al.: Constituição anotada da República Democrática de Timor-Leste. Escola de Direito da Universidade do Minho (2011). http://hdl.handle.net/10400.22/4008
28. Wu, D., He, D.: A study of query translation using google machine translation system. In: 2010 IEEE International Conference on Information Retrieval & Knowledge Management (CAMP), pp. 105–110. IEEE (2010). https://doi.org/10.1109/INFRKM.2010.5466911

29. Zhang, R., et al.: Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, 28July–2 August 2019, Volume 1: Long Papers, pp. 3173–3179. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/P19-1306
30. Zhou, D., Truran, M., Brailsford, T., Wade, V., Ashman, H.: Translation techniques in cross-language information retrieval. ACM Comput. Surv. (CSUR) **45**(1), 1–44 (2012). https://doi.org/10.1145/2379776.2379777